Towards Holistic, Pragmatic and Multimodal Conversational Systems

Pranava Madhyastha

City, University of London
pranava.madhyastha@city.ac.uk

Language acquisition and utilization transcend the mere exchange of lexical units. Visual cues, prosody, gestures, body movements, and context play an undeniably crucial role. Humans naturally communicate multimodally, employing multiple channels and synthesizing information from diverse modalities. My research delves into the characterization and construction of multimodal models that seamlessly integrate data from multiple independent modalities. I will cover recent work that highlights the challenges, achievements, and opportunities towards developing capable multimodal discursive models.

The first part of my talk will focus on understanding and exploring the potential contributions of the visual modality towards language tasks such as machine translation. I will present the noteworthy observation that machine translation models that incorporate visual contextual information perform more effectively when the available modalities are complementary rather than redundant (Caglayan et al. 2019). This finding aligns with the dominance effect in psychophysics, where it has been demonstrated that visual stimuli will consistently override auditory stimuli when the auditory information is purely complementary. I will then present some findings where we integrate visual information as a refinement process for generated text, enhancing and aligning the generated language with visual concepts (Ive, Madhyastha, and Specia 2019).

In the second part of my talk, I will delve into the significance of abstracting visual information to foster better alignment with linguistic inputs. Here, I will concentrate on systems that initially distill relevant visio-linguistic concepts and demonstrate that systems encompassing abstractions like object categories can aid in crafting a more interpretable vision and language system. This mitigates potential issues like hallucinations and constrains the system to be more anchored in the context of image captioning or visual question answering (Whitehouse, Weyde, and Madhyastha 2023). I will also explore some of the challenges and opportunities in translating these findings into multilingual settings and more complex phenomena such as toxic content prediction.

In the final segment of my talk, I will discuss recent work where we conducted a controlled study to scrutinize the impact of multiple modalities of information on the cognitive processing of language comprehension (Madhyastha, Zhang, and Vigliocco 2023). We performed experiments with human comprehenders using verbal stimuli in audio-only and audio-visual modalities. We juxtaposed the ERP signature (N400) associated with each word in audio-only and audiovisual presentations of the same verbal stimuli. We then assessed the extent to which surprisal measures (quantifying the predictability of words in their lexical context) were generated on the basis of distinct types of language models (specifically n-gram and Transformer models) that predicted N400 responses for each word. Our findings indicate that cognitive effort exhibits a substantial divergence between multimodal and unimodal settings. In addition, our findings suggest that while Transformer-based models, which have access to a broader lexical context, provide a superior fit in the audio-only setting, 2-gram language models (or language models that are only conditioned on local lexical context) are more effective in the multimodal setting. This highlights the considerable influence of local lexical context on cognitive processing in a multimodal environment.

These salient observations have a profound impact on the development of conversational systems, where interaction usually happens over multiple modalities and goes beyond lexical channels.

References

Caglayan, O.; Madhyastha, P.; Specia, L.; and Barrault, L. 2019. Probing the Need for Visual Context in Multimodal Machine Translation. In *NAACL 2023*.

Ive, J.; Madhyastha, P.; and Specia, L. 2019. Distilling Translations with Visual Awareness. In *ACL 2019*, 6525–6538. Florence, Italy.

Madhyastha, P.; Zhang, Y.; and Vigliocco, G. 2023. Are words equally surprising in audio and audio-visual comprehension? In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Whitehouse, C.; Weyde, T.; and Madhyastha, P. 2023. Towards a Unified Model for Generating Answers and Explanations in Visual Question Answering. In *EACL 2023*.